Gümüşhane University Journal of Science

GUFBD / *GUJS* (2023) 13(3): 767-779 doi: 10.17714/gumusfenbil.1253513 Research Article

SMOTE vs. KNNOR: An evaluation of oversampling techniques in machine learning

SMOTE ve KNNOR: Makine öğreniminde aşırı örnekleme tekniklerinin değerlendirilmesi

Ismet ABACI*[1], Kazim YILDIZ [2]

[1] *Marmara University, Institute of Pure and Applied Sciences, Computer Engineering, Istanbul*

[2]*Marmara University, Faculty of Technology, Computer Engineering, Istanbul*

• Received: 20.02.2023 • Accepted: 23.06.2023

Abstract

The increasing availability of big data has led to the development of applications that make human life easier. In order to process this data correctly, it is necessary to extract useful and valid information from large data warehouses through a knowledge discovery process in databases (KDD). Data mining is an important part of this, and it involves discovering data and developing models that extract unknown patterns. The quality of the data used in supervised machine learning algorithms plays a significant role in determining the success of predictions. One factor that improves the quality of data is a balanced dataset, where the input values are distributed close to each other. However, in practice, many datasets are unbalanced. To overcome this problem, oversampling techniques are used to generate synthetic data that is as close to real data as possible. In this study, we compared the performance of two oversampling techniques, SMOTE and KNNOR, on a variety of datasets using different machine learning algorithms. Our results showed that the use of SMOTE and KNNOR did not always improve the accuracy of the model. In fact, on many datasets, these techniques resulted in a decrease in accuracy. However, on certain datasets, both SMOTE and KNNOR were able to increase the accuracy of the model. Our results indicate that the effectiveness of oversampling techniques varies depending on the specific dataset and machine learning algorithm being used. Therefore, it is crucial to assess the effectiveness of these methods on a case-by-case basis to determine the best approach for a given dataset and algorithm.

Keywords: KNNOR, Machine learning, Oversampling, SMOTE, Unbalanced data

Öz

Büyük verinin artan mevcudiyeti, insan hayatını kolaylaştıran uygulamaların gelişmesine yol açmıştır. Bu veriyi doğru şekilde işlemek için, bilgi keşfi veritabanları (KDD) olarak adlandırılan büyük veri deposundan faydalı ve geçerli bilgiyi çıkarmak gereklidir. KDD işlemlerinin önemli bir parçası olan veri madenciliği, veriyi keşfetmeyi ve bilinmeyen desenleri çıkarmak için model geliştirmeyi içermektedir. Supervised makine öğrenimi algoritmalarında kullanılan verinin kalitesi, tahmin başarısının belirlenmesinde önemli bir rol oynar. Verinin kalitesini arttıran bir faktör, girdi değerlerinin birbirine yakın dağılmış olmasıdır. Ancak pratikte, birçok veri seti dengesizdir. Bu sorunu aşmak için, oversampling teknikleri gerçek veriye en yakın şekilde sentetik veri üretebilmek için kullanılır. Bu çalışmada, farklı veri setlerinde iki oversampling tekniği olan SMOTE ve KNNOR'un performanslarını farklı makine öğrenimi algoritmaları kullanarak karşılaştırdık. Sonuçlarımız, SMOTE ve KNNOR'un modellerin doğruluğunu her zaman arttırmadığını, hatta birçok veri setinde bu tekniklerin doğrulukta azalma yaratabileceğini gösterdi. Ancak belirli veri setlerinde, SMOTE ve KNNOR modellerin doğruluğunu arttırmayı başardı. Bulgularımız, oversampling tekniklerinin etkililiğinin belirli veri seti ve makine öğrenimi algoritmasına bağlı olarak değişebileceğini sugere etmektedir. Dolayısıyla, veri seti ve algoritma için en iyi yaklaşımı belirlemek için bu tekniklerin performanslarını durum bazında değerlendirmek önemlidir.

Anahtar kelimeler*: KNNOR, Makine öğrenmesi, Aşırı örnekleme, SMOTE, Dengesiz veri*

* Ismet ABACI; ismetabaci@marun.edu.tr

1. Introduction

With the advancement of technology, the proliferation of big data has made it possible to process and convert this data into applications that enhance human life. In order for the data to be processed correctly, first of all, it is necessary to be able to make sense of the data correctly. For this reason, it is necessary to determine patterns of useful, understandable, and valid data from large data warehouses (such as databases, data warehouses) a knowledge discovery process in database (KDD). Data mining, which is so crucial for the KDD process, is a process that discovers data and develops models that extract unknown patterns (Maimon & Rokach, 2010).

The quality of the data used in supervised machine learning algorithms has a very important effect on prediction success. Features such as the amount, accuracy, completeness, diversity, and balance of the data to be used are one of the requirements for supervised machine learning algorithms to give more accurate results.

One factor that improves the caliber of data used in machine learning algorithms is that the data is balanced (Galar et al., 2012). The distribution of the input numbers of the values to be estimated close to each other is called balanced, and the distribution of them far from each other is called unbalanced data sets. Models trained on unbalanced datasets often give poor results when they need to be generalized. In such cases, there are oversampling techniques used to get rid of the problems created by the unbalanced data set. By using these techniques, synthetic data can be produced as close to real data as possible. In this study, it will be examined which oversampling technique gives better results for which types of datasets and which machine learning algorithm. All the data set used in this study did not have an equal distribution, the data sets were balanced with the SMOTE and KNNOR oversampling techniques. Random Forest (RF), Decision Trees (DT), and Support Vector Machines (SVM) machine learning algorithms were used to compare how they affect prediction success.

Asif et al. (2015) developed a model for predicting fourth-grade university students' grades based on their social or demographic characteristics, which are unknown, but the first and second-grade course grades are known in their study. They applied Naive Bayes, 1-Nearest Neighbor, Decision Trees, and Neural Networks models to two different sets of data, containing grades from different time periods. While the 1-Nearest Neighbor model performed well in the first data set, they found that the Naive Bayes model was more successful in the second data set (Asif et al., 2015).

In 2005, two techniques named Borderline-SMOTE1 and Borderline-SMOTE2 were developed (Han et al., 2005). Both of these techniques focused on samples that were located on the boundary of the dataset, as the authors believed that these samples were at risk of being misclassified. By oversampling in this region, they aimed to decrease the chances of misclassification. Similarly, another technique called SAFESMOTE was developed which focused on oversampling the central samples that were farther away from the boundary region.

In their study, Adekitan & Salau (2019) aimed to predict the weighted grade point averages (GPAs) of university students in Nigeria using the GPAs of their first three years of education. They evaluated six different algorithms on the data, including Linear Regression, Random Forest, Decision Trees, Naive Bayes, Tree Trunk, and such as classifying it as a classification task and determining the success rate as a regression task. Neural Network. They found that the most successful model was the linear regression model, with an accuracy of 89.15% (Adekitan & O. Salau, 2019).

Strecht et al. (2015) conducted a study on determining the success or failure status of students and estimating their success grade. They considered determining the pass or fail status such as classifying it as a classification task and determining the success rate as a regression task. They found that while the model performed well in the classification task, the results were poor in the regression task. As a result, they found that SVM and DT had the best results in the classification task, while Adaboost, SVM and RF had the best results in the regression task (Strecht et al., 2015).

Recent studies in the field of SMOTE (Synthetic Minority Over-sampling Technique) have been concentrated on addressing the problem of uneven distribution of classes. One such algorithm is K-Means SMOTE (Last et al., 2017), that takes into account the spread of the under-represented group of samples in the data set. It

separates the data set into different clusters and calculates the proportion of imbalance in each cluster. The over-sampling is then done only on the samples where the proportion of imbalance is high.

Márquez et al. (2013) attempted to predict student performance using data from high school students in Mexico. They first reduced the dataset with 77 features to 15 features and applied the SMOTE method to balance the unbalanced dataset. After experimenting with various machine learning algorithms, they were able to achieve successful prediction results (Márquez et al., 2013).

Srinilta & Kanharattanachai (2021) balanced unbalanced datasets by using a search algorithm to to identify the ideal value for the k parameter in the SMOTE method. They then used a SVM for classification on six different datasets. They observed that the k value obtained by the search algorithm gave better results than the assumed k value (Srinilta & Kanharattanachai, 2021).

Flores et al. (2018) balanced a dataset used for sentiment analysis by applying the SMOTE method. They then evaluated the performance of the system by using Decision Vector Machine (DVM) and Naive Bayes (NB) algorithms on the balanced dataset. The study found that the 10-fold cross-validation method gave better results than the 0.7 hold-out method. Additionally, the study concludes that the SMOTE method improves the performance of both DVM and NB classifiers (Flores et al., 2018).

A review of related studies shows that machine learning methods such as Adaboost, NB, DT, SVM, and RF generally provide good results in predicting student success. Additionally, the application of SMOTE method to balance unbalanced data is found to improve the performance of these algorithms (Márquez et al., 2013; Srinilta & Kanharattanachai, 2021; Flores et al., 2018).

Douzas & Bacao (2018) introduced a new method for handling imbalanced datasets by employing Conditional Generative Adversarial Networks (cGANs) for effective data generation (Douzas & Bacao, 2018). Imbalanced datasets pose a significant challenge in machine learning, as the under-representation of minority classes can lead to biased predictions. The researchers compared the performance of cGANs with other oversampling techniques, including the widely-used Synthetic Minority Over-sampling Technique (SMOTE), on various datasets with different degrees of imbalance. They found that cGANs outperformed traditional oversampling methods, resulting in improved classification accuracy, particularly for minority classes. This study highlights the potential of using advanced generative models, such as cGANs, to tackle the issue of imbalanced datasets and enhance the performance of machine learning algorithms.

He & Garcia (2019) introduced the Adaptive Synthetic Sampling Approach (ADASYN) as a novel method for addressing imbalanced datasets in machine learning applications (He & Garcia, 2019). Imbalanced datasets present challenges when predicting minority classes, as these classes are often under-represented, leading to biased predictions. The researchers compared the performance of ADASYN with other popular oversampling techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), on various datasets with differing degrees of imbalance. They found that ADASYN provided more accurate classification results, particularly for minority classes, outperforming traditional oversampling methods. This study underscores the potential of using adaptive oversampling techniques, such as ADASYN, to tackle the issue of imbalanced datasets and improve the performance of machine learning algorithms in various applications.

Balci et al, (2022) demonstrated the potential of machine learning techniques in healthcare, specifically in detecting and classifying sleep-disordered breathing types using time and time-frequency features extracted from polysomnography records of 19 patients (Balci et al., 2022). The researchers processed six types of physiological data with digital signal processing methods to obtain 35 features, which were then subjected to various machine learning algorithms, including Artificial Neural Networks, Support Vector Machines, Random Forest, Naive Bayes, K-Nearest Neighbors, Decision Trees, and Logistic Regression. The study found that the Random Forest algorithm achieved the highest classification accuracy of 76.3% for the five-class scoring, which increased to 86.6% when Hypopnea was excluded. These findings underscore the value of machine learning algorithms in healthcare applications, particularly the distinctiveness of time and time-frequency domain features in sleep-disordered breathing scoring and pave the way for the development of diagnostic support systems capable of evaluating multiple polysomnography data concurrently.

In the literature, numerous studies have utilized machine learning algorithms for classifying agricultural products. Yasar (2023) used CNN models for bread wheat classification with an accuracy of 97.67%. Unlersen et al. (2022) employed a CNN-SVM hybrid model for wheat classification, achieving a 98.10% accuracy rate. Kaya & Saritas (2019) applied a neural network for durum wheat classification, reporting a 93.46% accuracy. Sabanci et al. (2022) combined CNN and SVM for pepper classification, obtaining a 99.02% accuracy rate. These studies demonstrate the potential of machine learning algorithms, including support vector machines and convolutional neural networks, for accurate agricultural product classification.

The application of machine learning algorithms in agriculture, particularly in the classification of various crop genotypes, has gained significant attention in recent years. Golcuk & Yasar (2023) conducted a groundbreaking study focusing on the classification of bread wheat genotypes using machine learning algorithms (Golcuk & Yasar, 2023). They collected 8,354 images from certified bread wheat varieties and extracted 90 color, 4 shape, and 12 morphological features using image processing and feature selection methods. These features were combined in various combinations and processed through an Artificial Bee Colony (ABC) algorithm for feature selection. The bread wheat genotypes were then classified using Support Vector Machines (SVM), Decision Tree (DT), and Quadratic Discriminant (QD) classifiers. To enhance the accuracy and objectivity of the classification process, the researchers performed a 10-fold cross-validation. The most successful classification process was achieved using the SVM algorithm, obtaining an accuracy rate of 96.28% with the 46 features selected by the ABC algorithm (Golcuk & Yasar, 2023). This study demonstrates the potential of machine learning algorithms in accurately classifying wheat genotypes, ultimately leading to more efficient harvests and higher income for farmers.

In a recent study, Islama et al. (2022) introduced the K-Nearest Neighbor Oversampling (KNNOR) approach as a solution for distinguishing vital and secure regions for oversampling and generating synthetic data points of the minority class. The KNNOR approach consists of three steps and takes into account the relative density of the entire population when generating artificial points. This method allows for more reliable oversampling of the minority class, as well as increased resistance to noise. It is a promising approach for addressing the issue of imbalanced datasets, where the minority class is underrepresented. This can be a common problem in machine learning and the KNNOR approach presents a valuable solution for identifying and utilizing the most effective oversampling regions.

2. Material and method

Datasets that have the same number of data in each class are called balanced datasets. There are two basic methods for balancing an unbalanced data set. One of them is oversampling, that is, increasing the quantity of samples in the minority class. The other is undersampling, that is, reducing the quantity of samples in the minority class (Islama et al., 2022). There are several approaches using these methods. One of these approaches is the SMOTE oversampling method.

2.1. Datasets

The study analyzed the performance of three machine learning algorithms in predicting outcomes on four datasets. The datasets include one related to secondary school student performance in Portugal, another related to hepatitis cases in India, and a third related to the reflectivity of land surface types as observed by the Landsat MSS satellite. The fourth dataset was not specified. Each dataset was evaluated using an unbalanced dataset and two different oversampling techniques, SMOTE and KNNOR, and the accuracy of the models was compared.

One of the datasets used in this study is the dataset used by Cortez and Silva in their studies, which includes the success of secondary school students in Portugal in Mathematics and Portuguese courses and various attributes that are predicted to affect this success (Cortez et al., 2008) The dataset includes demographic characteristics of students such as age and gender, social characteristics such as time spent going to school, weekly study hours, and attributes such as exam grades. The exam grades of the students are examined in three semesters, and the scores of the last semester are also accepted as the final assessment.

Again, according to the data set, exam grades were scored between 0 and 20. Since the records that do not contain data belonging to any of the attributes were not taken into account while preparing the data set, there was no need to perform a null value scan in this study.

Another dataset that is used in the study was the Hepatitis dataset on the UCI Machine Learning Repository website which is a collection of data on hepatitis cases in India (Gong et al., 1988). It contains 155 samples and 20 attributes, 16 of which are continuous and 4 of which are nominal. The target attribute for this dataset is "Class", which indicates whether a patient has hepatitis. The other 19 features are various characteristics of patients, such as their age, gender, and blood test results.

The hepatitis dataset includes the following characteristics: Class: Binary (presence or absence of hepatitis) Age: Continuous variable Gender: Binary (male or female) Steroid use: Binary (yes or no) Antiviral treatment: Binary (yes or no) Fatigue: Binary (yes or no) Malaise: Binary (yes or no) Loss of appetite: Binary (yes or no) Enlarged liver: Binary (yes or no) Liver function: Binary (yes or no) Detectable spleen: Binary (yes or no) Presence of spider angiomas: Binary (yes or no) Ascites: Binary (yes or no) Presence of esophageal varices: Binary (yes or no) Bilirubin levels: Continuous variable Alkaline phosphatase levels: Continuous variable SGOT levels: Continuous variable Albumin levels: Continuous variable Prothrombin time: Continuous variable Histological findings: Binary (yes or no). This dataset is commonly employed for diagnosing and categorizing hepatitis and serves as a benchmark for machine learning algorithms in relevant studies.

The last dataset was The Statlog (Landsat Satellite) dataset (Srinivasan, 1988) which is a collection of data on the reflectivity of certain land surface types as observed by the Landsat MSS satellite. Compiled by researchers at the University of California, Irvine and hosted in the UCI Machine Learning Repository.

The dataset consists of 6435 samples, each representing a piece of land on the Earth’s surface. For each sample, there are 36 properties, including various measures of reflectance at different wavelengths of light. There is also a class attribute, such as "barren land" or "evergreen forest", that indicates the type of land surface it represents, for example.

The dataset consists of 6435 samples, each representing a piece of land on the Earth’s surface. For each sample, there are 36 properties, including various measures of reflectance at different wavelengths of light. There is also a class attribute, such as "barren land" or "evergreen forest", that indicates the type of land surface it represents, for example.

We consider two tables to examine the class distributions and imbalance ratios of four distinct datasets shown in table 1 and table 2. The first table provides an overview of the datasets and their class distributions. Each row in the table represents dataset, and the columns show the number of instances for different classes (Attribute1 to Attribute6). For example, the Hepatitis dataset contains 123 instances of one class (Attribute1) and 32 instances of another class (Attribute2). To gain a better understanding of the degree of class imbalance, we computed imbalance ratios for each pair of classes in the datasets as shown in table 3. The imbalance ratio is calculated as the ratio of the number of instances in the majority class to the number of instances in the minority class. These calculated imbalance ratios provide insight into the distribution of instances between the datasets, with higher values indicating a more significant class imbalance.

Upon comparing the calculated imbalance ratios within each dataset shown in table 2, we observe that the Portuguese (student) dataset exhibits the highest degree of imbalance with an imbalance ratio of 5.49. This suggests that this dataset has a more disproportionate distribution of instances across its classes compared to the other datasets. Conversely, the Statlog dataset demonstrates the least imbalance, with an imbalance ratio of 2.45, indicating a relatively more even distribution of instances among its classes. By understanding the degree of imbalance within each dataset, practitioners can tailor their approaches accordingly to develop more accurate and robust classification models.

The mean absolute deviation, commonly referred to as the average distance, is a statistical measure that quantifies the dispersion of data points in a dataset relative to the mean. This metric, which is presented in Table 2, is calculated by first determining the mean of the dataset and then computing the absolute differences between each data point and the mean. The average of these absolute differences provides the mean absolute deviation, which serves as an informative indicator of the dataset's variability. In comparison to other measures

of dispersion, such as standard deviation, the mean absolute deviation is less sensitive to extreme values and offers a more straightforward interpretation of the average distance between data points and the mean. In our analysis, we calculated the mean absolute deviation to be 0.2925, which indicates that, on average, the data points deviate from the mean by approximately 0.2925 units. This measure provides insight into the dispersion of the data and can be helpful in understanding the overall distribution and consistency of the dataset.

Table 1. The distribution of unbalanced datasets.

Mineral	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
Hepatitis	123	32	-	-	-	-
Portuguese(student)	549	100	-	-	-	-
Mathematics(student)	130	103	62	60	40	-
Statlog	1533	1508	1358	707	703	626

Table 2. Dataset Characteristics and Imbalance Metrics

Dataset	Majority class	Minority class	Imbalance ratio	Standard deviation	Average distance
Hepatitis	123	32	3.84	0.40	0.33
Portuguese	549	100	5.49	0.17	0.13
Mathematics	130	40	3.25	0.23	0.17
Statlog	1533	626	2.45	0.37	0.32

Table 3. Imbalance Ratios Between Pairs of Datasets

Pair	Majority class	Minority class	Imbalance ratio
Hepatitis vs. portuguese	549	123	4.46
Hepatitis vs. mathematics	130	123	1.06
Hepatitis vs. statlog	1533	123	12.46
Portuguese vs. mathematics	549	130	4.22
Portuguese vs. statlog	1533	549	2.79
Mathematics vs. tatlog	1533	130	11.79

2.2. Machine learning algorithms

2.2.1. Support vector machines (SVM)

SVM is one of the prediction methods based on statistical learning frameworks (Cortes et al., 1995). It was proposed by Vapnik and Chervonenkis (1971). SVM aims to perform the classification task by drawing a hyperplane, which is a line in 2d or 3d. It tries to parse the previously marked data with this hyperplane it has drawn. While there can be multiple such hyperplanes, SVM tries to find the hyperplane that best separates the two categories. SVM is a supervised learning algorithm as it needs marked data.

2.2.2. Decision tree (DT)

Decision Trees are the classification of an unknown sample, which can be depicted using a tree diagram, by dividing it into sub-branches within the framework of certain rules. Decision trees consist of a root node, multiple internal nodes, and terminal nodes specifying the final classifications. A graph is obtained by calculating statistics for all classes, and decision limits are determined by looking at this graph. At each stage, the tree is created by selecting the attribute that best expresses the difference between the classes (Swain & Hauska, 1988).

2.2.3. Random forest (RF)

It was first put forward by Tin Kam Ho (1995). According to Breiman's definition, a Random Forest is a tree-structured collection of classifiers made up of randomly distributed vectors, augmented by voting for the most

popular class. The Random Forest algorithm is a method that does not cause excessive learning and has a successful prediction ability (Breiman, 2001).

2.3. Oversampling techniques

2.3.1. SMOTE

The SMOTE method, called the Synthetic Minority Oversampling Technique, is an over-learning technique that relies on creating artificial new samples instead of reproducing existing data. To multiply minority samples, a random sample is generated, with one sample between its k nearest neighbors. Considering a data point with position (6.4), let's assume that the position of its nearest neighbor is also (4.3). The coordinate of this data point is subtracted from the coordinate of its nearest neighbor ((4-6), (3-4)) and a new sample is created by multiplying the obtained value with a random value in the range 0-1. The newly created sample's position (x, y) is calculated as Equation 1 (Chawla et al., 2002).

$$(x, y) = (6,4) + rand(0,1) * (-2, -1) \quad (1)$$

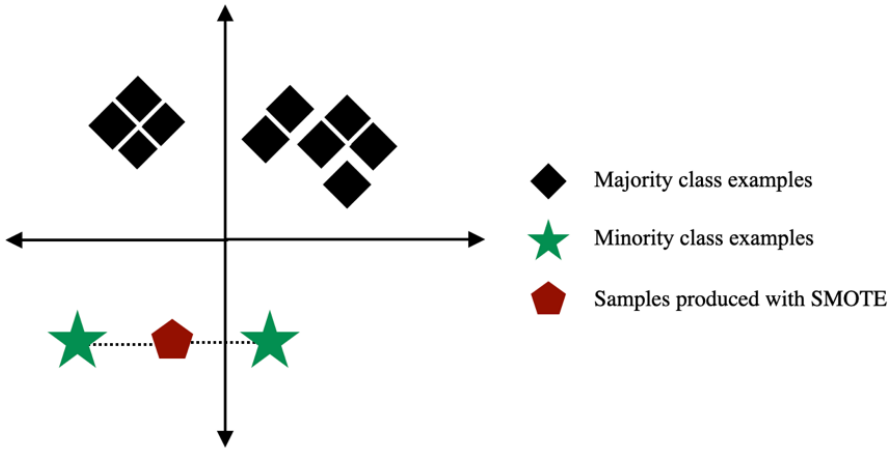

Figure 1. Generating synthetic data with SMOTE (Breiman, 2001)

In Figure 1, the data replication process is visualized by generating synthetic data at a random location between the minority samples and their k nearest neighbors with the SMOTE method.

2.3.2. KNNOR

KNNOR (K-Nearest Neighbor Oversampling) is a method for addressing class imbalance in a dataset by generating synthetic samples for the underrepresented class. Class imbalance can occur when one class in a dataset has significantly fewer examples than the other class, which can lead to biased machine learning models that are more likely to make predictions in favor of the more prevalent class.

One of the main advantages of KNNOR oversampling is that it does not necessitate any prior knowledge about the data distribution or the underlying relationships between the features and the target class. This makes it a relatively simple and flexible method that can be utilized for various datasets and classification tasks.

Overall, KNNOR oversampling can be a useful tool for addressing class imbalance in a dataset and enhancing the effectiveness of machine learning models. However, it is important to consider the potential drawbacks of this approach and to use it in combination with other techniques, such as feature selection or model hyperparameter tuning, to ensure optimal model performance.

3. Process flow

Before using mentioned oversampling techniques, Random Forest, Decision Trees, and Support Vector Machines machine learning algorithms were used to get the accuracy in order to compare with the data sets

which will be balanced using the oversampling techniques. The n-fold cross-validation method was applied in the training of the model and the results were noted. After this process, balancing the data process started.

It is a common mistake that the artificial data that was created in oversampling technique should not be used for testing the algorithm, it should be only used for training so that it wouldn't give inaccurate results. In order to achieve this technique the process flow shown in Figure 2 was used.

1. The data was divided into N random equal-sized subsamples.
2. One of the n subsamples which hadn't been used was separated.
3. First, the rest of the data were balanced with the oversampling methods.
4. The mentioned machine learning algorithms were trained with balanced data.
5. The test data was used to test the mentioned machine learning algorithms which were trained.
6. The results were noted.
7. If all the data was not tested, step one was repeated.
8. The overall results were calculated and noted to compare.

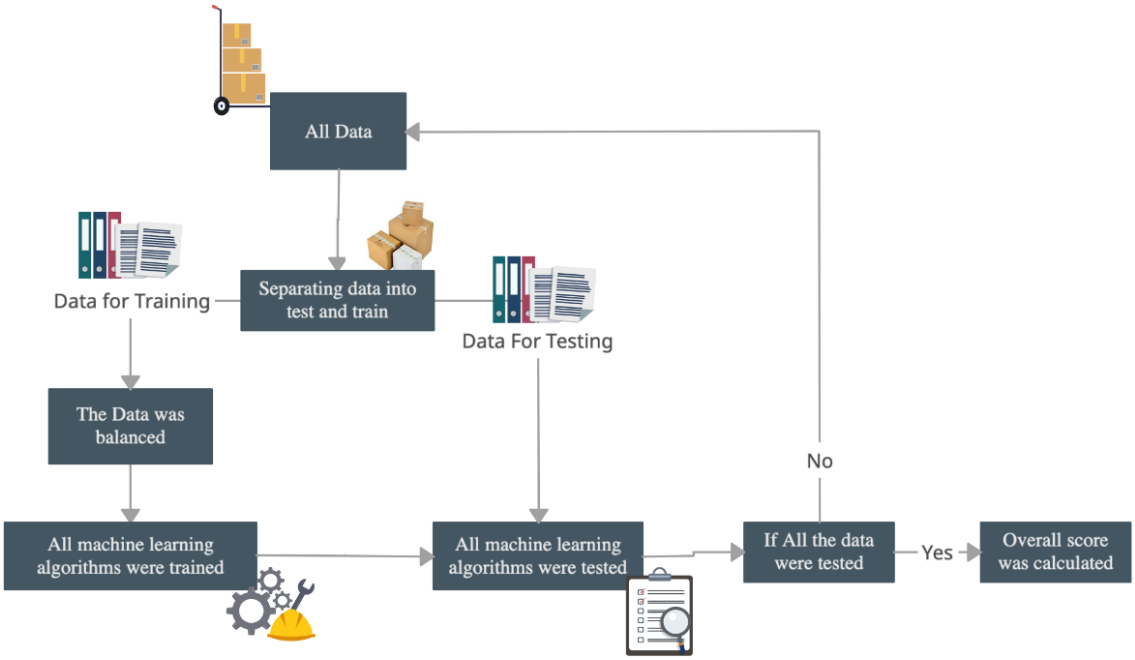

Figure 2. Architecture of proposed study

4. Results and discussion

We compared the performance of two oversampling techniques, SMOTE and KNNOR, on four different datasets using three machine learning algorithms: Decision Tree, Random Forest, and SVM. The datasets used in this study include the Hepatitis dataset, the Statlog dataset, the Mathematics dataset (Student), and the Portuguese dataset (Student). For each dataset, we evaluated the accuracy of the models using an unbalanced dataset and two different balanced datasets generated with SMOTE and KNNOR as shown in Table 4.

The units of data in Table 4 are expressed in terms of accuracy, which is a measure of how well a machine learning model performs. Accuracy is represented as a decimal value between 0 and 1, with 1 indicating perfect accuracy and 0 indicating the absence of accurate predictions. In this table, the accuracy values are presented with two decimal places, such as 0.77 or 0.91, which can also be interpreted as percentages by multiplying them by 100 (e.g., 77% or 91%).

Table 4. Accuracy results of the study with machine learning and high-speed sampling method.

Hepatitis dataset (2 classes)			
	Unbalanced data	**Balanced with smote**	**Balanced with knnor**
Decision tree	0.77	0.81	0.80
Random forest	0.87	0.85	0.85
SVM	0.79	0.57	0.76
The statlog dataset (6 classes)			
	Unbalanced data	**Balanced with smote**	**Balanced with knnor**
Decision tree	0.84	0.83	0.82
Random forest	0.90	0.90	0.90
SVM	0.88	0.87	0.87
Mathematics dataset (Student) (5 classes)			
	Unbalanced data	**Balanced with smote**	**Balanced with knnor**
Decision tree	0.66	0.68	0.67
Random forest	0.70	0.70	0.71
SVM	0.66	0.70	0.73
Portuguese dataset (Student) (2 classes)			
	Unbalanced data	**Balanced with smote**	**Balanced with knnor**
Decision tree	0.91	0.91	0.91
Random forest	0.93	0.93	0.93
SVM	0.91	0.92	0.92

Our results showed that the use of oversampling techniques did not always improve the accuracy of the model. In fact, in some cases, the accuracy decreased when using these techniques. The results also showed that the effectiveness of the oversampling techniques varied depending on the dataset and machine learning algorithm being used.

For the Hepatitis dataset, the Decision Tree model achieved an accuracy of 0.77 on the unbalanced dataset, 0.81 on the dataset balanced with SMOTE, and 0.80 on the dataset balanced with KNNOR. The Random Forest model achieved an accuracy of 0.87 on the unbalanced dataset, 0.85 on the dataset balanced with SMOTE, and 0.85 on the dataset balanced with KNNOR. The SVM model achieved an accuracy of 0.79 on the unbalanced dataset, 0.57 on the dataset balanced with SMOTE, and 0.76 on the dataset balanced with KNNOR.

For the Statlog dataset, the Decision Tree model achieved an accuracy of 0.84 on the unbalanced dataset, 0.83 on the dataset balanced with SMOTE, and 0.82 on the dataset balanced with KNNOR. The Random Forest model achieved an accuracy of 0.90 on all three datasets, while the SVM model achieved an accuracy of 0.88 on the unbalanced dataset, 0.87 on the dataset balanced with SMOTE, and 0.87 on the dataset balanced with KNNOR.

For the Mathematics dataset (Student), the Decision Tree model achieved an accuracy of 0.66 on the unbalanced dataset, 0.68 on the dataset balanced with SMOTE, and 0.67 on the dataset balanced with KNNOR. The Random Forest model achieved an accuracy of 0.70 on all three datasets, while the SVM model achieved an accuracy of 0.66 on the unbalanced dataset, 0.70 on the dataset balanced with SMOTE, and 0.73 on the dataset balanced with KNNOR.

For the Portuguese dataset (Student), the Decision Tree model achieved an accuracy of 0.91 on all three datasets. The Random Forest model achieved an accuracy of 0.93 on all three datasets, while the SVM model achieved an accuracy of 0.91 on the unbalanced dataset, 0.92 on the dataset balanced with SMOTE, and 0.92 on the dataset balanced with KNNOR.

It is important to note that the effectiveness of oversampling techniques is highly dependent on the nature of the data being analyzed. Our results indicate that oversampling techniques may not be necessary for datasets that are already well-balanced. In such cases, oversampling may actually introduce noise and lead to a decrease in accuracy.

Furthermore, our results suggest that the choice of machine learning algorithm is an important factor to consider when using oversampling techniques. For example, our findings showed that the performance of the SVM algorithm was significantly affected by the use of oversampling techniques. In contrast, the Random Forest algorithm was less sensitive to the use of oversampling techniques, and in some cases, showed no improvement in accuracy with the use of SMOTE or KNNOR.

The Portuguese (student) dataset exhibits the highest degree of imbalance with an imbalance ratio of 5.49, indicating a more disproportionate distribution of instances across its classes compared to other datasets. In contrast, the Statlog dataset has the least imbalance, with an imbalance ratio of 2.45, suggesting a relatively more even distribution of instances among its classes. Understanding the degree of imbalance within each dataset allows practitioners to tailor their approaches to develop more accurate and robust classification models accordingly.

We calculated the mean absolute deviation, also known as the average distance, to measure the dispersion of data points in a dataset relative to the mean. Our analysis showed that the mean absolute deviation was 0.2925, indicating that on average, data points deviate from the mean by approximately 0.2925 units. This measure provides insight into the dispersion of the data, which can be helpful in understanding the overall distribution and consistency of the dataset.

Our results revealed that the use of oversampling techniques did not consistently improve the accuracy of the model. In fact, in some instances, the accuracy decreased when using these techniques. The effectiveness of the oversampling techniques varied depending on the dataset and the machine learning algorithm being used. It is important to note that the effectiveness of oversampling techniques highly depends on the nature of the data being analyzed. Our results indicate that oversampling techniques may not be necessary for datasets that are already well-balanced, and in such cases, oversampling may introduce noise and lead to decreased accuracy.

After conducting a more detailed analysis, we discovered that the results obtained from these techniques varied depending on the specific machine learning algorithm employed. Specifically, we found that SMOTE gave a slightly better result when used with the Decision Tree algorithm, while KNNOR had a slightly better score when applied to SVM algorithms.

Overall, our study highlights the importance of carefully evaluating the effectiveness of oversampling techniques on a case-by-case basis. In some cases, oversampling techniques may not be necessary and may even lead to a decrease in accuracy. Therefore, it is crucial to carefully evaluate the performance of machine learning models with and without oversampling techniques in order to determine the best approach for a given dataset and algorithm. By doing so, we can ensure that the models we develop are accurate and reliable, and that we can extract useful and valid information from large datasets through the process of KDD and data mining.

Looking towards the future, there are several areas for further investigation. For example, additional research is needed to determine the optimal oversampling technique for the Hepatitis dataset with an SVM machine learning algorithm. It would also be useful to expand the scope of this study to include a wider range of datasets and machine learning algorithms, in order to have a more complete understanding of the performance of SMOTE and KNNOR. Additionally, it would be interesting to investigate the performance of these techniques in conjunction with other approaches for addressing class imbalance, such as undersampling or class weight adjustments.

Chawla et al. (2002) introduced the SMOTE algorithm, which demonstrated improved classification performance for minority classes and is applicable to various machine learning algorithms (Chawla et al., 2002). However, the effectiveness of SMOTE varies depending on the dataset and machine learning algorithm, and may introduce noise in some cases, leading to decreased accuracy. In our study, we found that the effectiveness of SMOTE and KNNOR oversampling techniques depends on the specific dataset and machine learning algorithm employed, which supports the notion that their performance can vary depending on the context.

Han et al. (2005) proposed the Borderline-SMOTE algorithm, which is a variation of the original SMOTE algorithm (Han et al., 2005). They demonstrated the effectiveness of their approach on several datasets, with improvements in classification performance over the original SMOTE. However, our study focused on comparing the original SMOTE and KNNOR techniques, and future research could consider evaluating Borderline-SMOTE as well. Since we did not evaluate Borderline-SMOTE in our study, we cannot directly compare its performance, but it may have potential similar limitations as the original SMOTE depending on the dataset and machine learning algorithm.

Batista et al. (2004) conducted a comprehensive study on various methods for balancing machine learning training data, including oversampling, under sampling, and hybrid approaches (Batista et al., 2004). Their findings highlighted the importance of using a suitable balancing method depending on the dataset and learning algorithm. Although their study does not specifically focus on SMOTE and KNNOR, it reinforces the idea that selecting the appropriate balancing method is crucial for improving classification performance.

Our study contributes to the existing body of knowledge by investigating the effectiveness of SMOTE and KNNOR oversampling techniques with different machine learning algorithms and datasets. We provided insights into the factors that influence the performance of oversampling techniques, such as the choice of machine learning algorithm and the nature of the dataset. However, we did not explore the performance of other oversampling techniques, such as Borderline-SMOTE, which limits our understanding of their potential benefits.

5. Conclusions

In conclusion, the study has shown that the performance of SMOTE and KNNOR oversampling techniques can vary significantly depending on the specific dataset and machine learning algorithm being used. Both methods did not consistently enhance the precision of the model, but were successful in raising the accuracy on specific datasets. The effectiveness of these methods varied depending on the type of machine learning algorithm applied, with SMOTE giving a slightly better result when used with the Decision Tree algorithm and KNNOR having a slightly better score when applied to SVM and Random Forest algorithms. These findings have important implications for the selection of oversampling techniques in machine learning applications and call for further research in this area. Overall, our results provide valuable insights into the performance of SMOTE and KNNOR oversampling techniques and can inform future research in this area.

Author contribution

For this study, the author responsible for the development of the methodology, writing the original draft, validation, and conceptualization is İsmet Abacı; the author responsible for the supervision, formal analysis, and writing is Kazim Yildiz.

Declaration of ethical code

The authors of this article declare that the materials and methods used in this study do not require ethical committee approval and/or legal-specific permission.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

Adekitan, A. I., & Salau, O. P. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, *5*(2), e01250. https://doi.org/10.1016/j.heliyon.2019.e01250

Ashwin Srinivasan (1988). UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting Student Academic Performance at Degree Level: A Case Study. *International Journal of Intelligent Systems and Applications*, *7*(1), 49–61. https://doi.org/10.5815/ijisa.2015.01.05

Balcı, M. A., Taşdemir, Ş., Ozmen, G., & Golcuk, A. (2022). Machine Learning-Based Detection of Sleep-Disordered Breathing Type Using Time and Time-Frequency Features. *Biomedical Signal Processing and Control*, *73*, 103402. https://doi.org/10.1016/j.bspc.2021.103402

Yasar, A. (2023). Benchmarking analysis of CNN models for bread wheat varieties. *European Food Research and Technology*, 249(3), 749-758.

Unlersen, M. F., Sonmez, M. M., Aslan, M. F., Demir, B., Aydin, N., Sabanci, K., & Ropelewska, E. (2022). CNN–SVM hybrid model for varietal classification of wheat based on bulk samples. *European Food Research and Technology*, *248*(8), 2043–2052. https://doi.org/10.1007/s00217-022-04029-4

Kaya, E., & Saritas, İ. (2019). Towards a real-time sorting system: Identification of vitreous durum wheat kernels using ANN based on their morphological, colour, wavelet and gaborlet features. Computers and Electronics in Agriculture, 166, 105016. doi:10.1016/j.compag.2019.105016

Sabanci, K., Aslan, M. F., Ropelewska, E., & Unlersen, M. F. (2022). A convolutional neural network-based comparative study for pepper seed classification: Analysis of selected deep features with support vector machine. Journal of *Food Process Engineering*, 45(6), e13955.

Batista, G. E. a. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. J., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Cortez, P., & Silva, A. L. (2008). Using data mining to predict secondary school student performance. *EUROSIS*. https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf

Douzas, G., & Bacao, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, 1–20. https://doi.org/10.1016/j.ins.2018.06.056

Douzas, G., & Bacao, F. (2018a). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, *91*, 464–471. https://doi.org/10.1016/j.eswa.2017.09.030

Flores, A. R., Icoy, R. I., Pena, C. L., & Gorro, K. D. (2018). *An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set*. https://doi.org/10.1109/iceast.2018.8434401

Gail Gong (1988). UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml

Galar, M., Fernández, A. Á., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man and Cybernetics*, *42*(4), 463–484. https://doi.org/10.1109/tsmcc.2011.2161285

Golcuk, A., & Yasar, A. (2023). Classification of bread wheat genotypes by machine learning algorithms. *Journal of Food Composition and Analysis*, *119*, 105253. https://doi.org/10.1016/j.jfca.2023.105253

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Lecture Notes in Computer Science* (pp. 878–887). Springer Science+Business Media. https://doi.org/10.1007/11538059_91

Islam, A., Samir, B. B., Rahman, A., & Bensmail, H. (2022). KNNOR: An oversampling technique for imbalanced datasets. *Applied Soft Computing*, *115*, 108288. https://doi.org/10.1016/j.asoc.2021.108288

Liu, J. (2021). Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data. *Soft Computing*, *26*(3), 1141–1163. https://doi.org/10.1007/s00500-021-06532-4

Maimon, O., & Rokach, L. (2009). Introduction to Knowledge Discovery and Data Mining. In *Springer eBooks* (pp. 1–15). https://doi.org/10.1007/978-0-387-09823-4_1

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, *38*(3), 315–330. https://doi.org/10.1007/s10489-012-0374-8

Srinilta, C., & Kanharattanachai, S. (2021). *Application of Natural Neighbor-based Algorithm on Oversampling SMOTE Algorithms*. https://doi.org/10.1109/iceast52143.2021.9426310

Strecht, P., Cruz, L. J., Soares, C. J., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. In *Educational Data Mining*. http://files.eric.ed.gov/fulltext/ED560769.pdf

Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, *15*(3), 142–147. https://doi.org/10.1109/tge.1977.6498972